A Database of Chemical Reactions Designed to Achieve Thermodynamic Consistency Automatically.

Peter M. May* and Kevin Murray[#]

AJ Parker Cooperative Research Centre in Hydrometallurgy, Murdoch University, Murdoch WA 6150, Australia

* Corresponding author. E-mail: may@murdoch.edu.au; Fax: +61-8-93101711.

\# Permanent address: Insight Modelling Services, P.O. Box 38953, Garsfontein East, Pretoria, South Africa 0060.

Abstract

The development of a large database, and associated computer software, for thermodynamic parameters of chemical reactions in aqueous solution is described. The aim has been to tackle some of the uncertainties that currently restrict the chemical modelling of aqueous, multicomponent mixtures.

A computer package, called JESS (Joint Expert Speciation System), is now available that can be informed progressively of new experimental data, critical data evaluations and established chemical trends so that sound and up-to-date thermodynamic calculations can be quickly and conveniently performed. This has been accomplished by storing a large amount of (inherently inconsistent) chemical data from the literature and, from this, producing a consistent set of mass balance equations by automatic means.

The basis of the chemical and numerical judgements made during this process is briefly outlined. The criteria for selection of best thermodynamic values and of best linear combinations of reactions are given particular attention.

Introduction

Serious discrepancies between published thermodynamic parameters of chemical reactions are well known (*e.g.* ref[1]). Since there are many different causes of these problems (such as experimental error, inadequate theory and carelessness), they can be very difficult to pinpoint and to eliminate. The situation is made worse because many thermodynamic data persisting in the scientific literature stem from values that are later corrected or become experimentally superceded.

In this regard, a great effort is needed for critical assessment of all the relevant primary measurements and for their transformation into thermodynamically consistent datasets. Such work not only demands considerable expertise and dedication but is also liable soon to be outdated as new measurements are made. This means that only few of the many tens of thousands of chemical reactions of interest have been adequately characterised.

To address these issues, we have developed a suite of computer programs that build and maintain a database of chemical reactions and their thermodynamic parameter values[2,3]. The database now has over 215,000 thermodynamic constants associated with more than 70,000 chemical reactions. Any form of chemical reaction can be handled, *e.g.* all the common types of solution equilibria including protonation, complex formation (binary and ternary), redox, solubility (of solids, liquids or gases) and adsorption. For each reaction, a practically unlimited number of associated thermodynamic values can be

stored. These reaction parameters can be equilibrium constants, Gibbs energies, enthalpies, entropies, heat capacities, or standard electrochemical potentials. Each such value is individually associated with (i) the conditions under which it was measured, *i.e.* the identity of any background electrolyte, the temperature, the pressure and the ionic strength, (ii) the reported uncertainty, (iii) the literature reference, (iv) an optional comment and (v) our current estimation of its reliability, on a weighting scale of 0 - 9.

This paper describes how these data are now being manipulated automatically to achieve thermodynamic consistency while simultaneously eliminating redundant reactions from equilibrium calculations. This has required the encoding of procedures to select data for quality and for relevance of experimental conditions as well as to establish the most appropriate linear combinations of reactions.

Justification

Achieving thermodynamic consistency is combinatorial in nature and, hence, can become prohibitive with large, multicomponent systems. Thorough (manual) evaluations are accordingly confined to small chemical assemblages (*e.g.*. ref.[4]) or can only be performed infrequently, as in the case of the more extensive NBS/NIST-sponsored projects (*e.g.*. ref.[5]). A computational procedure that automatically achieves thermodynamic consistency would evidently be of great assistance in terms of including the latest data,

reducing the burden of mundane work and minimizing errors. However, there are several other important, and perhaps more interesting, issues which can also be tackled by a chemically intelligent computer system. Two different kinds of problem, in particular, need to be overcome.

First, none of the other major compilations of equilibrium data for chemical speciation calculations[6-10] can be processed automatically. This is largely because they exist in a book-like format, be they electronic[6,7] or on paper[8-10], in which the reaction definition has not been recorded appropriately for numerical processing purposes. It is also because aqueous equilibrium data are often classified in terms of 'metals' and 'ligands', *i.e.* according to a pre-determined set of basis species, which involves implicit assumptions about kinetic constraints[11].

To avoid these limitations, an entirely general method of representing chemical reactions on the computer is necessary and has been developed. The format allows each reaction to be checked for charge and mass balance and, critically, it permits linear combination of reactions. Given the wide diversity of possible chemical relationships, it is important that there be no restriction of chemical reactions, nor of the chemical species involved in them, otherwise much available thermodynamic information can be lost. For example, neither the IUPAC[3] nor the NIST[4] electronic databases, which are the two largest individual collections of equilibrium constants, can ever be comprehensive, even for quite simple chemical systems, because they are not specifically designed to cope with

the stoichiometry of standard formation reactions, ternary complex formation or ligand-ligand interactions.

The second type of problem concerns computer modelling packages that have, hitherto, performed chemical speciation calculations using their own thermodynamic databases. Such databases, unfortunately, contain only a very small fraction of the available data (compared to the sources discussed above). However, of even greater concern, is the manner in which their thermodynamic data have been stored. Several fundamental issues, which are further addressed below, include (a) the storage of standard state values without the corresponding extrapolation functions and parameters used in their derivation, (b) the use of Gibbs energies of formation rather than the equilibrium constants, or Gibbs energies, of the most appropriate reactions, (c) the pre-determined, and hence invariant, choice of species whose concentrations or activities are treated as unknown - the so-called master or basis species - and (d) the failure to record properly why particular literature values have been selected in preference to others *and the full consequences of these decisions*.

The collective outcome of all these deficiencies has been to undermine the reliability of chemical equilibrium calculations to a considerable, and often surprising, extent. The striking differences in speciation distributions obtained by independent experts in a double-blind study commissioned by the European Nuclear Energy Agency was one of the major motivations for early JESS development[1]. There are still worryingly large

changes over time, both of the modelling predictions themselves (*e.g.* ref.[12]) and of the critically-assessed thermodynamic parameters on which such models are based (*e.g.* ref.[8]). There is little evidence to suggest that such changes are universally diminishing in magnitude as chemical systems become better characterised; on the contrary, the results of periodic critical evaluations of large systems of thermodynamic data, such as occurred in the CHEMVAL project[13], are often seemingly chaotic in nature (*e.g.* ref.[14], Tables 3 and 4 and ref[15], Figs. 6 and 8). It can take many years before even gross errors in modelling databases are detected (*e.g.* refs[16,17]). All this suggests that the root of the problem may be associated less with the arrival of new experimental values than it is with the way thermodynamic data are currently being processed.

Strategies

In our view, understanding the problems described above and, hence, finding a better way forward, hinges on a single proposition: full automation is essential to manipulate properly the sizeable and intricate datasets involved in modelling most aqueous chemical systems outside of the laboratory. This means bridging the gap between the large, book-like databases of IUPAC[6] and NIST[7] and the small, processable databases associated with modelling packages such as EQ3/6 , MINTEQA2 and SUPCRT92 (*e.g.* refs.[18-20]) that are currently being employed for most speciation modelling.

Two key aspects of this proposed linkage are (a) to make explicit the critical judgements and assumptions involved in selecting the thermodynamic data and (b) thereafter to achieve thermodynamic consistency by a fully automatic algorithm. In this way, the many decisions that always determine the results of large equilibrium models can be scrutinized and refined, so that hopefully outcomes will converge over time.

At present, it is practically impossible to trace all the thermodynamic parameters stored in modelling databases back to their origin in the chemical literature. Just as bad, comprehensive records of those data that have been rejected in the critical evaluation process are rarely, if ever, available. Accordingly, we generally store in the database all values that we locate in the literature but flag those considered 'worthless' to preclude their use and note reasons for the decision alongside rejected values. We also record all data in a form as close as possible to that of the original literature source so as to minimize transposition errors and to facilitate checking. This can be done because of the general way in which our reactions can be written and because redundant reactions are avoided automatically at a later stage.

One of the worst aspects of many thermodynamic compilations and databases for aqueous systems is the omission of details about the theoretical function(s), and parameters, adopted in the extrapolation to obtain infinite dilution (standard state) values. The contents of these databases thus tend to be regarded simplistically as 'right' or 'wrong', overlooking the complicated interplay between them and the chemical postulates

imbedded in any model, which together actually determine the models' quantitative predictions. It is of course only the correctness of these predictions that counts but, all too often, this is untested or achieved by parameterization.

It is not surprising that predictions fail when different functions or parameters are subsequently used with stored standard state values to back-calculate the chemical behaviour of real solutions. Discrepancies arising in this way are, nonetheless, often attributed to inadequacies in the theory of activity coefficient change! The obvious remedy is to ensure that the chemical information obtained experimentally is preserved in a complete and self-consistent way. This occurs, for instance, with sets of properly determined Pitzer or HKF coefficients. Our function is an extended version[21] of the Density model[22]. It has recently been shown to calculate satisfactorily the conditional equilibrium constants in aqueous solutions to high temperature (<250°C) and ionic strength (<5M)[21]. This function has the advantage that it minimizes the number of regressed parameters and, hence, we feel will improve and generalize our predictive capability.

Data Selection

Given that the chemical reactions, as stored in our database, can be manipulated algebraically, the mechanics by which thermodynamic consistency is achieved are

straightforward. All reactions of possible relevance to the model (as defined, for example, by the chemical elements in the system) are first extracted from the database; they are then converted, by a process of Gaussian elimination, into the consistent set of relationships that define the thermodynamic stability of all chemical species in terms of a basis set. These relationships appear as the linear combinations of reactions in the database that must be used to calculate the concentrations of species at equilibrium. The parameters of the extended Density model, which are periodically pre-determined for each reaction by regression of the raw values held in the database, are simply multiplied by the appropriate reaction coefficient (in the linear combination) and summed. Mass balance and other equations can then be constructed in the usual way and solved for any specified set of conditions.

The success of this approach is determined by the Gaussian elimination procedure. For the most part, this depends on how well each linear combination of reactions is selected. The optimum outcome is accomplished naturally by the Gaussian elimination provided that the reactions and species are processed in the correct order. The criteria for establishing this order are, unfortunately, equivocal. Broadly speaking, those reactions that are best characterised should be used first; this ensures that, if and when reactions are subsequently found to be redundant, those with the least 'information content' are the ones discarded. However, the optimum selection of species to be defined by each reaction (or set of reactions) depends on which species are predominant at equilibrium and, hence, can only be ascertained after the equilibrium calculation has been performed! Luckily, in

practice, most equilibrium calculations are not very sensitive to this choice and, in any event, the problem can be solved iteratively if necessary. On the other hand, it is very useful to be able to change the basis set easily: all databases with pre-determined basis / master species are prone to difficulties associated with mathematical ill-conditioning when the concentrations of these species, as the algebraic unknowns, become too small.

As now implemented in JESS, the ordering prior to Gaussian elimination is based on 'scores' which are assigned to both species and reactions, according to their respective criteria. Species are favoured as unknowns (in the ultimate equilibrium calculation) if they are involved in many well-characterised reactions. Reactions are selected by assessing the number and proximity of their equilibrium constants occurring under conditions of ionic strength and temperature closest to those required by the model. In this way, different linear combinations of reactions are employed for equilibrium calculations at different conditions. For example, an overall formation constant (β_2) may be preferred over the combination of stepwise constants (K_1K_2) when it has been better characterised *under the appropriate conditions* than the second association constant (K_2), and *vice versa*.

The 'scores' for each reaction also depend on the assessed quality of the data. This is indicated by the 'weights' associated with each thermodynamic parameter, as described above. For example, one important criterion is whether there has been only a single value

determined for a reaction: multiple values measured independently under the same conditions that agree are given greater weights.

The weights thus provide a key mechanism by which the system is informed of chemical judgements and opinion. They quantify assessments made not only by us, the compilers of the database, but also as appear in the literature in the form of critical reviews or expert comment. When the weights are modified to reflect a changing view of particular thermodynamic constants, the chemical knowledge embodied in the database evolves. In this way the database may progressively become both comprehensive and critical.

Conclusion

Judgements of data and selection of reactions are always involved in the development of thermodynamic databases for chemical equilibrium modelling. With the present approach, it is now possible for all of a model's primary data to be documented, for opinions to be kept explicit and for the decision-making process to be codified. This means that equilibrium models can be constructed more transparently, more objectively and more reproducibly. In particular, errors of fact, of judgement and of procedure now all ought to be detected more easily and, hence, corrected more quickly.

Literature Cited

(1)   Grauer, R. Solubility Limitations: An "Old Timer's" View, in*: Modelling in Aquatic Chemistry*, Grenthe, I.; Puigdomench, I. (Eds.), Nuclear Energy Agency, OECD, Paris, France, 1997, pp. 131-152.

(2)   May, P. M.; Murray, K. JESS, A Joint Expert Speciation System - I. Raison d'Etre. *Talanta*, **1991**, 38, 1409-1417.

(3)   May, P. M.; Murray, K. JESS, A Joint Expert Speciation System - II. The Thermodynamic Database. *Talanta*, **1991**, 38, 1419-1426.

(4)   Königsberger, E., Königsberger, L-C., Gamsjäger, H. Low-temperature Thermodynamic Model for the System Na_2CO_3-$MgCO_3$-$CaCO_3$-H_2O. *Geochim. Cosmochim. Acta*, **1999**, 63, 3105-3119.

(5)   Chase, M. W. Jr., NIST-JANAF Thermochemical Tables. 4th Edition, *Journal of Physical and Chemical Reference, Monograph 9*, 1998.

(6)   Pettit, L. D.; Powell, H. K. J. *IUPAC Stability Constants Database*, Academic Software, Otley, U.K., 1993

(7)   Martell, A. E.; Smith, R. M.; Motekaitis, R.J. *Critically Selected Stability Constants of Metal Complexes. Version 5.0; Standard Reference Database 46*, NIST, Gaithersburg, MD, U.S.A, 1998.

(8)   Barin I.; Platzki G. *Thermochemical Data of Pure Substances*, 3rd Edn., Weinheim, Germany, 1995.

(9)   Sillén, L. G.; Martell, A. E. *Stability Constants of Metal-ion Complexes*, Chem. Soc. London, Spec. Publ. 17, 1964 and Spec. Publ. 25, 1971.

(10)   Lide D.R.; Frederikse H. P. R. *CRC Handbook of Chemistry and Physics, 76th Edn.*, CRC Press, Cleveland, Ohio, U.S.A., 1995.

(11)   May, P. M.; Murray, K. New Approaches for Modelling the Chemistry of Complicated Environmental Systems, *219th Book of Abstracts, Part 1, ACS National Meeting, San Francisco,* March 26-30, 2000, Paper ENVR 109.

(12)   May, P. M. Modelling Metal-Ligand Equilibria in Blood Plasma, in: *Handbook of Metal-Ligand Interactions in Biological Fluids, 1. Bioinorganic Chemistry. Vol. 2*, Berthon, G. (Ed.), Marcel Dekker, NY, U.S.A., 1995, pp. 1184-1194.

(13)   Read, D.; Broyd, T.W. Recent Progress in Testing Chemical Equilibrium Models: The CHEMVAL Project. *Radiochim. Acta*, **1991**, 52, 453-456.

(14)   Tardy, Y.; Nahon, D. Geochemistry of Laterites, Stability of Al-Goethite, Al-Hematite, and Fe^{3+}-Kaolinite in Bauxites and Ferricretes. *Am. J. Sci.*., **1985**, 285, 865-903.

(15)   Arnórsson, S.; Stefánsson, A. Assessment of Feldspar Solubility Constants in Water in the Range 0° to 350°C at Vapor Saturation Pressures. *Am. J. Sci.*, **1999**, 299, 173-209.

(16)   Onysko, S. J.; McNearny, R. L. GIBBTEQ: A MINTEQA2 Thermodynamic Data Error Detection Program. *Ground Water*, **1997**, 35, 912-914.

(17)   Serkiz S. M.; Allison J. D.; Perdue, E. M.; Allen, H. E.; Brown, D.S. Correcting Errors in the Thermodynamic Database for the Equilibrium Speciation Model MINTEQA2. *Wat. Res.*, **1996**, 30, 1930-1933.

(18)   Wolery, T. J.; Jackson, K. J.; Bourcier, W. L.; Bruton, C. J.; Viani, B. E.; Knauss, K.G.; Delany, J. M. Current Status of the EQ3/6 Software Package for Geochemical Modeling. *Am. Chem. Soc. Symp. Ser. 416*, 1990, pp. 104-116.

(19)   Allison, J.D.; Brown, D. S.; and Novo-Gradac. *MINTEQA2/PRODEFA2, A Geochemical Assessment Model for Environmental Systems: Version 3.0 User's Manual*. U.S. Environmental Protection Agency, Athens, U.S.A., EPA/600/3-91/021, 1991.

(20)   Johnson, J. M.; Oelkers, E. H.; Helgeson, H. C. SUPCRT92: A Software Package for Calculating the Standard Molal Thermodynamic Properties of Minerals, Gases, Aqueous Species and Reactions from 1 to 5000 bars and 0 to 1000C. *Comput. Geosci.*, **1992**, 18, 899-947.

(21)   May, P. M. A Simple, General and Robust Function for Equilibria in Aqueous Solutions to High Ionic Strength and Temperature. *Chem. Comm.*, **2000** (submitted).

(22)   Anderson, G.M.; Castet, S.; Schott, J.; Mesmer, R. E. The Density Model for Estimation of Thermodynamic Parameters of Reactions at High Temperatures and Pressures. *Geochim. Cosmochim. Acta*, **1991**, 55, 1769-1779.